MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2017/2018

TDS3551 – DATA MANAGEMENT

(All Sections / Groups)

16 OCTOBER 2017
9:00 a.m. – 11:00 a.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 6 pages (including cover page) with 5 Questions only.

2. Attempt **ALL** questions. The distribution of the marks for each question is given. This paper carries **100 marks.**

3. Please print all your answers in the Answer Booklet provided in this examination.

Question 1 (20 Marks)

a) In a traditional Relational DataBase Management System (RDBMS), a batch queue can be used to address the voluminous writes to the database which would otherwise overwhelm the RDBMS. Explain, with the help of a diagram, how this technique works to alleviate the issue.

[4 marks]

b) Explain what is meant by sharding of a database?

[2 marks]

c) Provide at least 2 (TWO) examples with brief descriptions for each of the following sources of data:

i. Databases and Files
ii. Web & Crowd Sourced Data
iii. Social Media
iv. Internet of Things (IoT)

[8 marks]

d) State the functionality of the following tools:

i. Yarn
ii. Sqoop
iii. Pig
iv. Flume
v. HBase
vi. Kafka

[6 marks]

Question 2 (10 Marks)

a) State and briefly explain at least 6 (SIX) desired properties of a Big Data Architecture.

[12 marks]

Continued ...

b) Illustrate, using a diagram, the inter-relationship between the 3 major components of Big Data Architecture which comprises of the Batch Layer, the Serving Layer and the Speed Layer. Explain the flow of the data from the incoming (and historical) data to the queries made by the applications.

[6 marks]

c) Batch processing provides for high-throughput but generally is also a high latency process. Explain the two terms, throughput and latency.

[2 marks]

Question 3 (10 Marks)

a) Using the database table below, develop a fact-based model representation for the following data. State any assumptions or rationale behind your model.

ID	Name	Date of Birth	Gender	Employer	Location	Date Stamp
701010102333	Dave	10-Oct-70	Male	IBM	New York	30/6/2017
830410142565	Bob	10-Apr-83	Male	HP	Bangkok	3/7/2017
880613149982	Anne	13-Jun-88	Female	Oracle	Manila	3/7/2017
910130045279	Ivan	30-Jan-91	Male	Teradata	Kuala Lumpur	4/7/2017
920523068844	Susan	23-May-93	Female	Teradata	Manila	4/7/2017
701010102333	Dave	10-Oct-70	Male	IBM	Singapore	6/7/2017
701010102333	Dave	10-Oct-70	Male	Microsoft	Singapore	3/9/2017
910130045279	Ivan	30-Jan-91	Male	Teradata	Singapore	30/9/2017

[5 marks]

b) For the GraphML code below, illustrate it graphically. (In other words, draw the graph).

```
<graphml>
 <graph>
  <node id="1"/>
  <node id="2"/>
  <node id="3"/>
  <node id="4"/>
  <edge id="e1" source="1" target="2"/>
  <edge id="e2" source="2" target="3"/>
  <edge id="e3" source="2" target="4"/>
  <edge id="e4" source="3" target="1"/>
 </graph>
</graphml>
```

[4 marks]

Continued ...

c) The Master Dataset is generally too large to exist on a single server and is continuously growing at a rapid pace. A file system seems to be an appropriate way to store this data that should have immutable property. Explain what is meant by immutable in this context.

[3 marks]

d) The Hadoop Distributed File System (HDFS) has two categories of nodes; the DataNode and the NameNode. Give a brief explanation for each category.

[2 marks]

e) In computing on the master dataset, using incremental algorithm seems to be more efficient than re-computing on the whole master dataset. However, there are some trade-offs. State 3 (THREE) trade-offs with comparisons between the two methods (incremental and re-computation).

[6 marks]

Question 4 (10 Marks)

a) When ingesting data into the batch layer to produce the batch views, some data normalization is normally required. Explain what does this normalization process do and give an example.

[4 marks]

b) However, normalized data has lower performance (computational speed) than using de-normalized data. Explain, using examples if necessary, on why this is so?

[2 marks]

c) The serving layer has 4 (FOUR) main requirements, which are Batch Writes, Scalability, Random Reads and Fault Tolerant. This makes the management and maintenance simpler as there isn't a need for Random Writes. Random Writes are the main cause of database complexity. Explain 2 (TWO) functions that a Random Write system is required to have that increases the complexity of the system.

[4 marks]

Continued ...

d) One of the main challenges of the Serving Layer is to serve the Batch Views in a scalable manner. In achieving the scalability, the data can be spread out across many database instances.

URL	Bucket	Pageviews
foo.com/blog/1	0	10
foo.com/blog/1	1	21
foo.com/blog/1	2	7
foo.com/blog/1	3	38
foo.com/blog/1	4	29
bar.com/post/a	0	178
bar.com/post/a	1	91
bar.com/post/a	2	568

a. If we are querying for a batch view that returns a smallish result (in kB), by having the data spread out, the latency to obtain from all the various servers will negate the benefits of distributing them. Explain why the latency will be an issue using the sample data above. The data relates to the number of URL pageviews per hour (which is recorded as [Bucket] in the dataset). You can expand on it if necessary for better clarity.

[4 marks]

b. How would you propose to solve it (using the dataset explained above as an example) and what are the potential drawbacks of your solution?

[6 marks]

Question 5 (10 Marks)

a) What is the purpose of the Speed Layer and why is it not addressed by the Batch Layer and the Serving Layer?

[4 marks]

b) One of the requirements of a Big Data Architecture is Fault Tolerant. In the Speed Layer, it operates similarly to a RDBMS except that the data is transient. Explain how does this effect the human fault tolerance of the overall system.

[4 marks]

Continued ...

c) The Speed Layer stores data for a few minutes or hours. The Speed Layer produces realtime views that represent recent data and is a continuously updated at a rapid pace. Assuming that the overall system receives about 1.5 TeraByte of data a day where data arrives in an average size of 1 MegaByte. The Batch Layer is scheduled to generate Batch Views 6 times a day. How many times does the Speed Layer needs to do the computation (assuming re-computation method) and on average, what is the size of each of the computation?

[6 marks]

d) What does CAP (in CAP Theorem) stands for and briefly explain the CAP Theorem.

[2 marks]

e) Expiring the data from the Speed Layer is dependent on the Batch Layer computation and the availability of the batch views on the Serving Layer. Explain why two sets of realtime views are used to ensure that the expiring of the data from the Speed Layer will not cause issues.

[4 marks]

End of Page